

# The Role of Data Cap in Optimal Two-part Network Pricing

Xin Wang<sup>†</sup>, Richard T. B. Ma<sup>\*</sup>, Yinlong Xu<sup>†</sup>

<sup>†</sup> School of Computer Science and Technology, University of Science and Technology of China

<sup>\*</sup> School of Computing, National University of Singapore

yixinxa@mail.ustc.edu.cn, tbma@comp.nus.edu.sg, ylxu@ustc.edu.cn

**Abstract**—Internet services are traditionally priced at flat rates; however, many Internet service providers (ISPs) have recently shifted towards two-part tariffs where a data cap is imposed to restrain data demand from heavy users. Although the two-part tariff could generally increase the revenue for ISPs and has been supported by the US FCC, the role of data cap and its optimal pricing structures are not well understood. In this article, we study the impact of data cap on the optimal two-part pricing schemes for congestion-prone service markets. We model users' demand and preferences over pricing and congestion alternatives and derive the market share and congestion of service providers under a market equilibrium. Based on the equilibrium model, we characterize the two-part structures of the revenue- and welfare-optimal pricing schemes. Our results reveal that 1) the data cap provides a mechanism for ISPs to transition from the flat-rate to pay-as-you-go type of schemes, 2) both of revenue and welfare objectives move the pricing towards usage-based schemes with diminishing data caps, and 3) the welfare-optimal tariff comprises lower fees than the revenue-optimal counterpart, suggesting that regulators might want to promote usage-based pricing but regulate the lump-sum and per-unit fees.

## I. INTRODUCTION

Traditionally, Internet service providers (ISPs) use flat-rate pricing [4] for network services, where users pay fixed monthly fees for unlimited data usage. Flat-rate pricing was widely adopted because it was easy for ISPs to implement and was preferred by users for its simplicity. However, with the rapid development and growing popularity of data intensity services, e.g., online video streaming and cloud-based applications, the Internet traffic keeps growing more than 50% per annum [5], which exposes some disadvantages of the flat-rate scheme. Because flat-rate does not count the users' data usage, "bandwidth hogs" [6] consume an unfair share of capacity and are subsidized by normal users, and ISPs cannot generate enough revenues to recoup their costs, especially for the mobile providers. As a consequence, most mobile LTE providers [7] and even broadband ISPs, e.g., Verizon [8] and AT&T [9], start to introduce a data cap and adopt a two-part tariff structure, a combination of flat-rate and usage-based pricing. Under such a two-part scheme, additional charges are

imposed if a user's data demand exceeds the data cap and the exceeded amount is charged based on a per-unit fee.

Although prior work [6, 10–12] has shown that data-capped schemes could help ISPs generate higher revenues than that under the flat-rate pricing and the FCC chairman has recently backed usage-based pricing for broadband to penalize heavy Internet users [13], little is known about 1) the data cap's role and impact on the revenue-optimal two-part pricing, and 2) potential regulations on two-part pricing for protecting social welfare from monopoly providers.

In this paper, we focus on a generic congestion-prone service market, e.g., mobile, broadband or cloud services, and study the data cap under two-part pricing schemes. Unlike physical commodities, the quality of network service is intricately influenced by a negative network effect (or network externality): the more users access the service simultaneously, the worse performance it provides. We model this service congestion as a function of providers' capacity and their data load. We characterize users by their demand and values on data usage, and analyze the market shares and congestion levels of the providers under varying pricing and market structures. Based on our model, we analyze the effect of data cap on the provider's optimal revenue and pricing structure. We also analyze and compare the revenue-optimal and welfare-optimal two-part schemes, and derive regulatory implications. Our main contributions and findings include the following.

- We model users' optimal data usages and preferences over various pricing schemes and exogenous levels of congestion (Section III). We characterize the existence and uniqueness (Theorem 1) of a market equilibrium and show its monotonic dynamics (Theorem 2) under varying pricing and market structures.
- We analyze the impact of data cap on a provider's optimal revenue and pricing decision under two-part tariff (Section IV). We find that diminishing the data cap transforms revenue-optimal pricing from flat-rate to pay-as-you-go type of schemes, during which the provider's revenue changes from a minimum to a maximum (Theorem 3).
- We characterize the dynamics of welfare-optimal two-part tariff (Section V). We find that the optimal social welfare varies from a minimum to a maximum with decreasing data cap (Theorem 4) and the welfare-optimal pricing imposes lower fees than the revenue-optimal counterpart. The results suggest that regulators might want to encour-

The updated version serves an erratum to prior papers [1–3], which studied the role of data cap in optimal two-part tariffs. In Section VI and V of [1] and [3], and Section III and VI of [2], the theoretical analysis and simulation results of the optimal pricing do not hold in general. In this present article, we combine the models from the original papers with new corrected results in Section IV and V.

age the use of two-part pricing with limited data cap, while regulate the lump-sum and per-unit fees.

We believe that our work provides new insights into the role of data cap in the optimal structure of two-part tariff. Our results could help service providers design revenue-optimal pricing schemes and guide regulatory authorities to legislate desirable regulations.

## II. RELATED WORK

Early studies of two-part tariff came from economics. Oi [14] first studied the price discrimination via quantity discounts in the monopoly Disneyland market. Calem *et al.* [15] examined and compared the revenue-optimal prices by a multi-product monopoly and a differentiated oligopoly. Littlechild [16] studied the explicit characterizations of welfare-optimal pricing and the effect of consumption externalities. Scotchmer [17] explored the nature of Nash equilibrium among profit-maximizing shared facilities. However, all of these work were confined to a special case where the data/usage cap is set to be zero. Our work studies the impact and dynamic of data cap on the two-part pricing, which generalizes the special case.

As the demarcation between the lump-sum and usage-based fees, data cap plays a crucial role in the two-part tariff structures. Prior work [6, 10–12, 18] demonstrated that data-capped schemes can increase providers' revenue compared with the traditional flat-rate pricing [4]. Odlyzko *et al.* [19] showed that ISPs could also reduce network congestion by imposing data caps. Our analysis and results also confirm these observations. Furthermore, we focus on understanding the impact of data cap on the structures of revenue-optimal and welfare-optimal tariffs. The impact and optimal design of data cap have been empirically studied. In a qualitative study of households users under bandwidth caps, Chetty *et al.* [20] studied how the uncertainties of user types and demand would impact the setting of data cap and operator's revenue, and proposed new tools to help users manage their caps. Poularakis *et al.* [21] proposed a framework to calculate the optimal data caps and empirically evaluated the gains of ISPs when they adopt data caps based on traffic datasets. Unlike these efforts, our work adopts an analytical approach to characterize the desirable data cap and the optimal structure of the two-part pricing schemes.

More generally, there have been several works that study the usage-based Internet pricing. Hande *et al.* [22] characterized the economic loss due to ISPs' inability or unwillingness to price broadband access based on the time of use. Li *et al.* [23] studied the optimal price differentiation under complete and incomplete information. Basar *et al.* [24] devised a revenue-maximizing pricing under varying user scale and network capacity. Shen *et al.* [25] investigated optimal nonlinear pricing policy design for a monopolistic service provider and showed that the introduction of nonlinear pricing provides a large profit improvement over linear pricing. In this paper, we focus on the two-part pricing. Besides optimizing the revenue from the provider's perspective, we also look into the welfare-optimal solution, through which we derive regulatory implications.

From a modeling perspective, Chander [26], Reitman [27], Ma [28] and our work all consider the service market with congestion externalities. Chander [26] studied the quality differentiation strategy of a monopoly provider and Reitman [27] studied a multi-provider price competition. Both of them modeled the market as a continuum of non-atomic users, each of which is characterized by a quality-sensitivity parameter. However, this one-dimensional model only applies for flat-rate pricing and the distribution of users was often assumed to be uniform for analytical tractability. To faithfully characterize the utility of users under two-part tiered pricing, we establish a novel two-dimensional model that describes users by their data demand and valuation on data usage. Furthermore, we analyze a class of distributions, including the uniform distribution, to understand the impact of user demand and value on the optimal pricing structures of the providers. Ma [28] also considered a two-dimensional user model; however, the author only focused on the pay-as-you-go pricing, a special case of the two-part tariff structure studied in this paper.

## III. MODEL

### A. Model of Users and Their Data Demand

We model each user by two orthogonal characteristics: her average value of per-unit data usage  $v$  and desirable data demand  $u$ . The user's data usage is measured by what she is billed, e.g., the number of bits transmitted or the amount of time being online.

We denote  $q$  as the congestion level of an ISP. Given the network congestion  $q$ , we denote  $\rho(u, q)$  as the user's achievable demand.

**Assumption 1:**  $\rho(u, q): \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$  is a continuous function, increasing in  $u$  and decreasing in  $q$ . It has an upper-bound  $\rho(u, 0) = u$  and satisfies  $\lim_{q \rightarrow +\infty} \rho(u, q) = 0$ .

Assumption 1 states that a user's achievable data demand equals its desirable demand  $u$  under no congestion and decreases monotonically when the network congestion  $q$  becomes more severe.

Beside network congestion, the user's actual data usage also depends on her ISP's pricing. We consider an ISP that adopts a two-part tiered pricing structure  $\theta = (g, f, p)$ , where  $g, f$  and  $p$  denote a data cap, a lump-sum service fee and a per-unit usage fee, respectively. Under this scheme, we denote  $t(y, \theta)$  as the user's charge when  $y$  units of data are consumed, defined as

$$t(y, \theta) \triangleq f + p(y - g)^+.$$

Intuitively, if a user's usage is below the data cap  $g$ , the ISP only collects the lump-sum fee  $f$ ; otherwise, extra charges are imposed on the usage above the data cap with the per-unit usage fee of  $p$ . This two-part structure is also a generalization of flat-rate, i.e.,  $g = +\infty$ , and pay-as-you-go [28] pricing, i.e.,  $f = 0$  and  $g = 0$ .

We denote  $\pi(y)$  as the user's utility when she consumes  $y$  units of data, defined by  $\pi(y) \triangleq vy - t(y, \theta)$ , i.e., the total traffic value minus the charge. We assume that users determine

their optimal data usages that maximize their utilities. In other words, each user tries to solve the following problem:

$$\begin{aligned} & \text{Maximize} \quad \pi(y) = vy - t(y, \theta) \\ & \text{subject to} \quad 0 \leq y \leq \rho(u, q), \end{aligned} \quad (1)$$

where a user's actual data usage is constrained by the achievable demand  $\rho(u, q)$  under the network congestion  $q$ . We define  $\phi = (u, v)$  as the type of the user and denote  $y^*(\phi, \theta, q)$  as its optimal usage under ISP's pricing scheme  $\theta$  and congestion  $q$ . By solving the optimization problem in (1), we derive the unique optimal solution for any user type  $\phi$  as

$$y^*(\phi, \theta, q) = \rho(u, q) - [\rho(u, q) - g]^+ \mathbf{1}_{\{v < p\}}. \quad (2)$$

Intuitively, if a user's achievable demand  $\rho(u, q)$  is beyond the data cap  $g$  and her value  $v$  is lower than the per-unit usage fee  $p$ , she will avoid consuming extra usage above  $g$  which would result in reducing her utility, and thus the optimal data demand equals the data cap, i.e.,  $y^* = g$ ; otherwise, the optimal data demand equals her achievable demand, i.e.,  $y^* = \rho(u, q)$ . Equation (2) also shows that the optimal demand will increase if the value  $v$  or desirable demand  $u$  increases, or the provider's per-unit fee  $p$  or data cap  $g$  or congestion  $q$  alleviates.

### B. Users' Preferences over Providers

We consider a market that comprises of a set  $\mathcal{N}$  of providers. We denote  $\theta = (\theta_i : i \in \mathcal{N})$  and  $\mathbf{q} = (q_i : i \in \mathcal{N})$  as the pricing strategy and congestion vectors of the providers. We define  $y_i^*(\phi) \triangleq y^*(\phi, \theta_i, q_i)$  as the optimal demand of user type  $\phi$  when it chooses provider  $i$ . For any two providers  $i, j \in \mathcal{N}$ , we denote  $i \succ_\phi j$  if users of type  $\phi$  prefer  $i$  over  $j$ .

**Definition 1:** We denote  $\pi_i(y)$  as the user's utility function when using provider  $i$ . For any  $i, j \in \mathcal{N}$ ,  $i \succ_\phi j$  if and only if 1)  $\pi_i(y_i^*(\phi)) > \pi_j(y_j^*(\phi))$  or 2)  $\pi_i(y_i^*(\phi)) = \pi_j(y_j^*(\phi))$  and  $i$  is chosen over  $j$  by the user based on any arbitrary tie-breaking condition.

Based on Definition 1, we assume that users of type  $\phi$  would choose their favorite provider  $i$  that induces the highest utility under their optimal data usage, i.e.,  $i \succ_\phi j, \forall j \in \mathcal{N} \setminus \{i\}$ . However, a user's best provider might still induce negative utility, and therefore we allow users not to use any of the providers if they all induce negative utility as follows.

**Assumption 2:** There exists a provider, indexed by 0, that sets  $f_0 = p_0 = 0$  and maintains a fixed level of congestion  $q_0$ .

Assumption 2 states that there always exist a free provider that users can choose. When we set the congestion level  $q_0 = +\infty$ , the free provider becomes a dummy provider which users could choose so as to guarantee zero utility. Besides, when the congestion  $q_0 < +\infty$ , the free provider can be used to characterize lower-end competitors in the market that have capacities to accommodate users with a certain congestion quality  $q_0$ . Because in practice, users might also find public options [29] such as free municipal WiFi and municipal broadband [30] recently supported by the US FCC.

To distinguish between the free provider and normal providers, we assume that the set  $\mathcal{N}$  are all normal providers, i.e.,  $0 \notin \mathcal{N}$ .

Based on Definition 1 and Assumption 2, we denote  $\Phi_i$  as provider  $i$ 's market share, i.e., the set of user types that choose to use  $i$ , defined by

$$\Phi_i(\theta, \mathbf{q}, q_0) = \{\phi : i \succ_\phi j, \forall j \in \mathcal{N} \cup \{0\} \setminus \{i\}\}. \quad (3)$$

Next, we show how providers' market shares  $\Phi_i(\theta, \mathbf{q}, q_0)$  vary when the set of competing providers  $\mathcal{N}$  or the pricing decisions  $\theta$  change.

**Proposition 1:** For a set  $\mathcal{N}$  of providers, if two pricing decisions  $\hat{\theta}$  and  $\theta$  satisfy  $\hat{g}_i \geq g_i, \hat{f}_i \leq f_i, \hat{p}_i \leq p_i$  for some  $i \in \mathcal{N}$  and  $\hat{\theta}_j = \theta_j$  for all  $j \neq i$ , we have  $\Phi_i(\theta, \mathbf{q}, q_0) \subseteq \Phi_i(\hat{\theta}, \mathbf{q}, q_0)$  and  $\Phi_j(\theta, \mathbf{q}, q_0) \supseteq \Phi_j(\hat{\theta}, \mathbf{q}, q_0), \forall j \neq i$ . For two sets  $\mathcal{N}$  and  $\mathcal{N}'$  of providers, if  $\mathcal{N} \subseteq \mathcal{N}'$  and  $(\theta'_i, q'_i) = (\theta_i, q_i)$  for  $\forall i \in \mathcal{N}$ , we have  $\Phi_i(\theta', \mathbf{q}', q_0) \subseteq \Phi_i(\theta, \mathbf{q}, q_0), \forall i \in \mathcal{N}$ .

Proposition 1 states that under fixed levels of congestion  $\mathbf{q}$ , the market share of a provider would increase if the provider reduces its fees  $f$  or  $p$ , or raises its data cap  $g$ , unilaterally. Meanwhile, the market share of any other provider will decrease. It implies that monopolistic providers could use fees and data cap to trade off its market share and revenue; while, oligopolistic providers could compete for market shares by decreasing their fees and increasing data caps. Proposition 1 also implies that bringing new providers into the market will intensify market competition and existing providers' market shares will decrease, because some of their users might switch to the new providers.

Proposition 1 holds under the condition of fixed network congestion. However, a provider's congestion level depends on its market share and the data usage of its users. We discuss the dynamics of network congestion and equilibrium in the next subsection.

### C. Network Congestion and Equilibrium

We denote  $U$  and  $V$  as the maximum desirable data demand and maximum per-unit data value of all users. Thus, the domain of users is defined as  $\Phi = [0, U] \times [0, V]$ . We model the set of all users by the measure space  $(\Phi, \mu)$ , where  $\mu$  denotes a product measure

$$\mu(E_1 \times E_2) = \mu_u(E_1) \times \mu_v(E_2), \quad \forall E_1 \subseteq [0, U], E_2 \subseteq [0, V],$$

where  $\mu_u$  and  $\mu_v$  are two continuous measures, defined by

$$\begin{aligned} \mu_u((u_1, u_2]) &= F_u(u_2) - F_u(u_1), \forall u_1 \leq u_2, \text{ and} \\ \mu_v((v_1, v_2]) &= F_v(v_2) - F_v(v_1), \forall v_1 \leq v_2, \end{aligned}$$

for some non-decreasing distribution functions  $F_u$  and  $F_v$ .

Based on the distribution of users, we denote  $d_i$  as provider  $i$ 's data load, i.e., the aggregate data demand of users of  $i$ , defined by

$$d_i = D(\Phi_i(\theta, \mathbf{q}, q_0); \theta_i, q_i) \triangleq \int_{\Phi_i(\theta, \mathbf{q}, q_0)} y_i^*(\phi, \theta_i, q_i) d\mu \quad (4)$$

On the one hand, given congestion levels  $\mathbf{q}$ , provider  $i$  has an induced data load  $d_i = D(\Phi_i; \theta_i, q_i)$ . On the other hand, the provider's congestion level  $q_i$  is influenced by its data load  $d_i$ .

We denote  $c_i$  as provider  $i$ 's capacity and model its congestion as a function  $q_i = Q_i(d_i, c_i)$  of data load  $d_i$  and capacity  $c_i$ .

**Assumption 3:**  $Q_i(d_i, c_i) : \mathbb{R}_+^2 \mapsto \mathbb{R}_+$  is continuous, increasing in  $d_i$ , decreasing in  $c_i$  and satisfies  $Q_i(0, c_i) = 0$ .

Different forms of the congestion function  $Q_i$  can be used to model the different technologies used by the provider. Assumption 3 implies that a provider  $i$ 's congestion increases (decreases) when its data load  $d_i$  (capacity  $c_i$ ) increases, and no congestion exists when no user consumes data from the provider.

We denote  $Q_i^{-1}(q_i, c_i)$  as the inverse function of  $Q_i(d_i, c_i)$  with respect to  $d_i$ , which defines the implied load under the capacity  $c_i$  and an observed congestion level of  $q_i$ . By Assumption 3, we know that  $Q_i^{-1}(q_i, c_i)$  is continuous, increasing in both  $q_i$  and  $c_i$ , and satisfies  $Q_i^{-1}(0, c_i) = 0$ . We denote  $\mathbf{c} = (c_i : i \in \mathcal{N})$  as the vector of the capacities of all the providers. Under an exogenous level  $q_0$  of congestion and when the providers make exogenous pricing decisions  $\theta$  and capacity planning decisions  $\mathbf{c}$ , the resulting congestion  $\mathbf{q}$  of the providers can be determined endogenously when users choose their best providers. We define such a market equilibrium of the system as follows.

**Definition 2:** Given the congestion level  $q_0$ , for a set  $\mathcal{N}$  of providers with any fixed pricing strategies  $\theta$  and capacities  $\mathbf{c}$ ,  $\mathbf{q}$  is an equilibrium if and only if

$$q_i = Q_i\left(D(\Phi_i(\theta, \mathbf{q}, q_0); \theta_i, q_i), c_i\right), \quad \forall i \in \mathcal{N}.$$

To better understand the above definition, we can equivalently rephrase the above equality condition as  $D(\Phi_i(\theta, \mathbf{q}, q_0); \theta_i, q_i) = Q_i^{-1}(q_i, c_i)$ , where the left-hand side is the induced data load of provider  $i$  given its market share  $\Phi_i(\theta, \mathbf{q}, q_0)$ , pricing strategy  $\theta_i$  and congestion  $q_i$  and the right-hand side is its implied data load under capacity  $c_i$ . In equilibrium, both equal the actual aggregate user demand  $d_i$ . Because the equilibrium is depend on the pricing strategies  $\theta$ , capacities  $\mathbf{c}$  and congestion  $q_0$ , we also denote it as  $\mathbf{q}(\theta, \mathbf{c}, q_0)$ .

**Theorem 1:** Under Assumption 1-3, for any fixed pricing strategies  $\theta$ , capacities  $\mathbf{c}$  and congestion  $q_0$ , there always exists a market equilibrium  $\mathbf{q}$ , satisfying  $q_i < q_0, \forall i \in \mathcal{N}$ . In particular, when the market has only one normal provider, i.e.,  $|\mathcal{N}| = 1$ , the equilibrium is unique.

Theorem 1 states that under minor assumptions of the demand (Assumption 1) and congestion (Assumption 3), the existence of a market equilibrium can be guaranteed. Besides, the normal providers always have lower congestion levels than  $q_0$  of the free provider under equilibrium; otherwise, no user would use the normal providers. When the market consists only one normal provider, the equilibrium is unique. For this case, we denote  $I$  as the provider and define  $\mathcal{I} = \{I\}$ . The next theorem shows how its congestion level varies when its pricing strategy  $\theta_I$  and capacity  $c_I$  change, or other providers enter the market to compete.

**Theorem 2:** In a market  $\mathcal{I}$ , provider  $I$ 's unique level of congestion  $q_I(\theta_I, c_I, q_0)$  in equilibrium is non-increasing in its fees  $f_I$ ,  $p_I$  and capacity  $c_I$ , but is non-decreasing in its

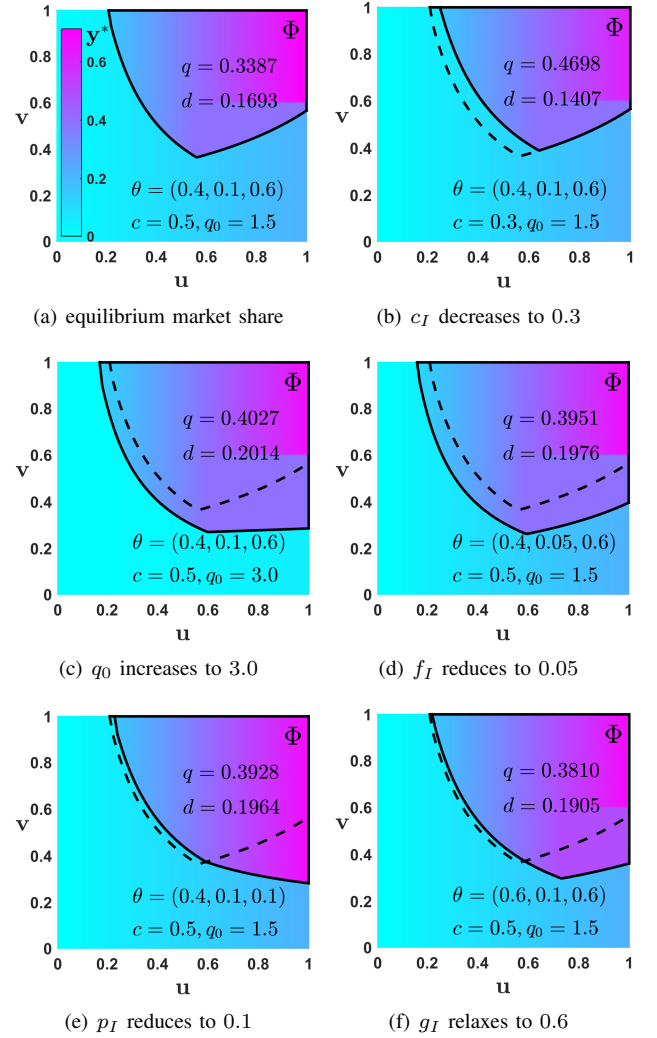


Fig. 1. Shift of market share for the provider

data cap  $g_I$  and congestion  $q_0$  of the free provider. When new providers enter and form a new market  $\mathcal{N} \supset \mathcal{I}$  and  $I$  keeps its pricing  $\theta_I$  and capacity  $c_I$ ,  $I$ 's congestion  $q_I(\theta, \mathbf{c}, q_0)$  under any new equilibrium satisfies  $q_I(\theta, \mathbf{c}, q_0) \leq q_I(\theta_I, c_I, q_0)$ .

Theorem 2 states that when the normal provider increases fees or decreases data cap, its induced congestion level will be alleviated, because its market share as well as data load will decrease. When the provider increases capacity, the congestion level will also decrease, although the resulting market share and data load will increase. Furthermore, if the congestion  $q_0$  deteriorates, the provider's congestion will increase, as some users of the free provider will shift to it. It also shows that with more competing providers, the existing monopoly's congestion will decrease, as users have more choices and may switch to other providers. This implies that competition could alleviate the congestion at the providers, because effectively more providers bring higher capacity to the entire market.

Under any market equilibrium  $\mathbf{q}(\theta, \mathbf{c}, q_0)$ , we denote  $R_i$  and  $S_i$  as the revenue and social welfare generated by provider  $i$ ,

defined as

$$R_i \triangleq \int_{\Phi_i} t(y_i^*(\phi), \theta_i) d\mu \quad \text{and} \quad S_i \triangleq \int_{\Phi_i} v y_i^*(\phi) d\mu, \quad (5)$$

where  $y_i^*(\phi) = y^*(\phi, \theta_i, q_i)$  and  $\Phi_i = \Phi_i(\theta, \mathbf{q}, q_0)$  are evaluated at the equilibrium  $\mathbf{q}(\theta, \mathbf{c}, q_0)$ .

#### D. Model Parameters and Properties

Although our model is built upon generic assumptions (Assumption 1 and 3), it does not yet capture the characteristics of network services. To this end, we carefully choose the model parameters, i.e., the congestion function  $Q_i(d_i, c_i)$ , the achievable demand function  $\rho(u, q)$  and the measure space  $(\Phi, \mu)$  of user domain.

Next, we choose a quintessential form  $\rho(u, q) = ue^{-q}$  for the achievable demand function, which is used by prior work [27, 31]. Under this form, the user's achievable demand decays exponentially at a rate of  $q$ , i.e., the level of congestion. We adopt the congestion function  $Q_i(d_i, c_i) = d_i/c_i$ , which models the *capacity sharing* [32] nature of network services. This form has been used in much prior work such as [32–34]. **Corollary 1:** Suppose  $\rho(u, q) = ue^{-q}$  and  $Q_i(d_i, c_i) = d_i/c_i$  for all  $i \in \mathcal{N}$  and let  $\mathbf{q}$  be an equilibrium under parameters  $\theta, \mathbf{c}, q_0, \Phi$  and  $\mu$ . For another market with  $\hat{\mathbf{f}} = \mathbf{f}/(UV)$ ,  $\hat{\mathbf{g}} = \mathbf{g}/U$ ,  $\hat{\mathbf{p}} = \mathbf{p}/V$ ,  $\hat{\mathbf{c}} = k\mathbf{c}/U$ ,  $\hat{q}_0 = q_0$ ,  $\hat{\Phi} = [0, 1] \times [0, 1]$  and  $\hat{\mu}([0, \hat{u}] \times [0, \hat{v}]) = k\mu([0, U\hat{u}] \times [0, V\hat{v}])$  for all  $(\hat{u}, \hat{v}) \in \hat{\Phi}$  and any  $k > 0$ , we must have  $\hat{\mathbf{q}} = \mathbf{q}$  as an equilibrium under which  $\hat{d}_i = kd_i/U$  for all  $i \in \mathcal{N}$ .

Corollary 1 states that under the exponential form of achievable demand and linear congestion function, the measure and domain of users can be normalized to be a probability measure and  $[0, 1] \times [0, 1]$ , respectively. In particular, keeping the congestion level of the free provider unchanged, the equilibrium does not change when 1) the providers' capacities and the user size scale linearly at the same rate  $k$ , 2) the lump-sum fees, data caps, capacities and the users' desirable demands are normalized by  $U$ , and 3) the lump-sum and per-unit fees and users' values are normalized by  $V$ .

Based on this result, we can focus on probability distribution functions  $F_u(U) = F_v(V) = 1$  and the maximum user demand and valuation  $U = V = 1$  without loss of generality. Further, we will consider the forms  $F_u(x) = x^\alpha$  and  $F_v(x) = x^\beta$  for  $x \in [0, 1]$ , where  $\alpha$  and  $\beta$  model the distribution of users with respect to their desirable data demands and values, respectively. For instance, when  $\beta = 1$ , user values are uniformly distributed; otherwise, they are leaning toward the high ( $\beta > 1$ ) or low ( $\beta < 1$ ) values in the domain  $[0, 1]$ . In summary, we will analyze the two-part tiered pricing of the providers with the following assumption.

**Assumption 4:** Any provider  $i$ 's congestion satisfies  $Q_i(d_i, c_i) = d_i/c_i$ , the users are distributed by  $F_u(x) = x^\alpha$ ,  $F_v(x) = x^\beta$  for  $x \in [0, 1]$ , and their achievable demands satisfy  $\rho(u, q) = ue^{-q}$ .

When the level of congestion is exogenously given, Proposition 1 implies that the market share of a provider will decrease

when it raises fees, e.g.,  $f_I$  and  $p_I$ , or reduces data cap  $g_I$ . However, the higher fees or lower data cap would alleviate the provider's congestion in equilibrium by Theorem 2, which results in attracting more congestion-sensitive users to join. As a result, the dynamics of the provider's market share combines both effects and may not be monotonic as we illustrate through an example in Figure 1 under Assumption 4. In this example, the users are uniformly distributed, i.e.,  $\alpha = \beta = 1$ . In each subfigure, x-axis and y-axis vary the desirable data demand  $u$  and value  $v$  of the user types. In other words, each point in the subfigures represents a unique user type. We use different colors of points to represent the data usages  $y^*$  of user types. In subfigure (a), the upper-right area shows the market share  $\Phi_I$  of the normal provider when it makes the pricing strategy  $\theta_I = (g_I, f_I, p_I) = (0.4, 0.1, 0.6)$  and capacity  $c_I = 0.5$ , under the congestion  $q_0 = 1.5$ . These parameters induce congestion  $q_I = 0.3387$  and data load  $d_I = 0.1693$  in equilibrium. From subfigure (a) to (b), the provider's capacity  $c_I$  is reduced from 0.5 to 0.3, which exacerbates the congestion level and results in a smaller market share for the provider. From subfigure (a) to (c), the market is less competitive, i.e.,  $q_0$  increases from 1.5 to 3.0; from subfigure (a) to (d), the provider uses a lower lump-sum fee, i.e.,  $f$  decreases from 0.1 to 0.05. Both cases lead to a larger market share  $\Phi_I$  as well as higher congestion  $q_I$  and data load  $d$  of the provider. From subfigure (a) to (e) and (f), the provider decreases the per-unit fee to  $p_I = 0.1$  and increases the data cap to  $g_I = 0.6$ , respectively. The lower per-unit fee and higher data cap also induce higher data load  $d_I$  and congestion  $q_I$  for the provider; however, the resulting market share  $\Phi_I$  attracts more low-value heavy users and loses some high-value light users. In all cases, the congestion of the provider in equilibrium is always smaller than the congestion  $q_0$  of market competitors which is consistent with the result of Theorem 1.

#### IV. REVENUE-OPTIMAL PRICING

We studied the market shares, congestion levels and data loads of providers under a market equilibrium in the previous section. In this section, we further explore how providers set the two-part pricing strategy so as to optimize their revenues. Because the data cap is the demarcation between the lump-sum (for demand below the data cap) and usage-based (for demand above the data cap) charges, we focus on the impact of data cap on the provider's optimal revenue and pricing decisions and identify the role of data cap in the two-part pricing structure.

We start with a unique normal provider in the market and use the free provider with congestion  $q_0$  to model its lower-end competitors. Under any given data cap  $g_I$ , we assume the ISP chooses the lump-sum and per-unit fees, i.e.,  $f_I$  and  $p_I$ , to maximize its revenue  $R_I$ , which is formulated by the following optimization problem.

$$\begin{aligned} & \text{Maximize} && R_I(g_I, f_I, p_I), \\ & \text{subject to} && f_I, p_I \geq 0. \end{aligned} \quad (6)$$

We denote  $f_I^*(g_I)$  and  $p_I^*(g_I)$  as an optimal solution of (6), and  $R_I^*(g_I)$  as the corresponding maximum revenue. Next,

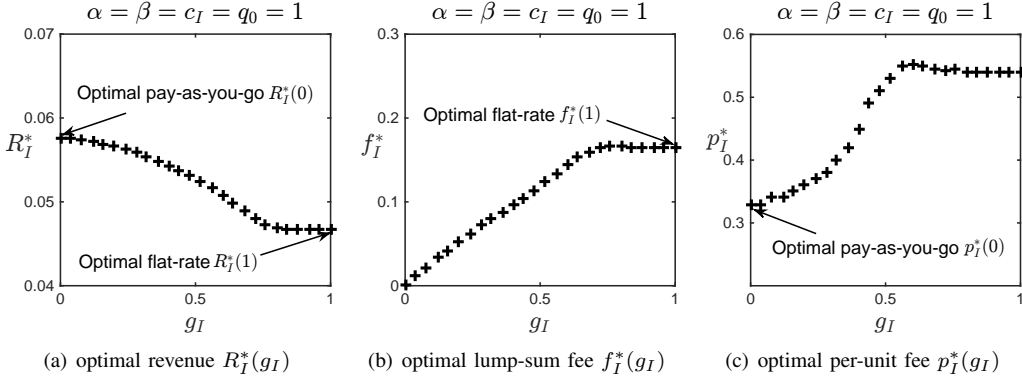


Fig. 2. Maximum revenue, optimal lump-sum and per-unit fee under varying data cap.

we study how various data caps influence its optimal revenue  $R_I^*$ , and corresponding pricing decisions, i.e.,  $f_I^*$  and  $p_I^*$ . In particular, we compare the two-part pricing with the pay-as-you-go scheme, i.e.,  $f_I = g_I = 0$ , and the flat-rate scheme, i.e.,  $g_I = +\infty$ . For simplification, we define the maximum revenue under the flat-rate as  $R_\infty \triangleq \lim_{g_I \rightarrow \infty} R_I^*(g_I)$ .

We consider the models under Assumption 4 where the maximum desirable demand is normalized to  $U = 1$ . As a result, an effective data cap  $g_I$  that might influence the users' demand has to be less than or equal to 1. In other words, when  $g_I$  is larger than 1, the two-part pricing is equivalent to a flat-rate scheme, i.e.,  $R_I^*(g_I) = R_\infty = R_I^*(1)$  for any  $g_I \geq 1$ . Thus, we will focus on  $g_I \in [0, 1]$  without loss of generality. Figure 2 shows how the data cap  $g_I$  influences the ISP's maximum revenue  $R_I^*(g_I)$ , and the optimal fees  $f_I^*(g_I)$ ,  $p_I^*(g_I)$ , under the uniform user distribution, i.e.,  $\alpha = \beta = 1$ , and normalized ISP capacity and free provider's congestion, i.e.,  $c_I = q_0 = 1$ . Subfigures (a) to (c) plots  $R_I^*$ ,  $f_I^*$  and  $p_I^*$  as functions of the data cap  $g_I$  varying along the x-axis, respectively.

From Figure 2(a), we observe that the optimal revenue  $R_I^*$  decreases monotonically with the data cap  $g_I$ . Specifically, 1) when the pricing converges to the flat-rate structure as  $g_I$  goes to one, the optimal revenue reaches a minimum; and 2) when the pricing converges to the (pure) usage-based structure as  $g_I$  tends to zero, the optimal revenue gets a maximum. From Figures 2(b) and 2(c), we see that the optimal fees  $f_I^*$ ,  $p_I^*$  both show increasing trends as the data cap  $g_I$  is relaxed. In particular, as the data cap is zero, the lump-sum fee is also zero, implying that the optimal two-part pricing is a pay-as-you-go scheme. The following theorem shows that our observations on the optimal revenue and fees are not particular to the model parameters, i.e.,  $\alpha = \beta = c_I = q_0 = 1$ .

**Theorem 3:** For any congestion  $q_0$ , parameters  $\alpha, \beta, c_I > 0$  under Assumption 4, the provider's optimal revenue satisfies

$$R_\infty \leq R_I^*(g_I) \leq R_I^*(0) \text{ for } \forall g_I \geq 0.$$

Besides, when the data cap is zero, the revenue-optimal lump-sum fee is zero, i.e.,  $f_I^*(0) = 0$ .

Theorem 3 shows that the revenue generated under an

optimal two-part pricing with any fixed data cap  $g_I (\geq 0)$  is no less than that under an optimal flat-rate scheme. In other words, an ISP could always be better off by switching from an optimal flat-rate scheme to an optimal two-part pricing scheme. Intuitively, under any fixed lump-sum  $f_I$ , the two-part structure generalizes the flat-rate scheme under which  $p_I = 0$  and the data cap does not play a role. Consequently, even without changing its flat-rate  $f_I$ , the ISP could restrict  $g_I$  and increase  $p_I$  to trade off between higher usage revenue from high-value users' demand and its market share, which potentially lead to higher total revenue. This result is consistent with the views in prior work [10–12] that data cap could help providers extract higher revenue from the market.

Theorem 3 also states that when the data cap is zero, the revenue from the optimal two-part pricing reaches a maximum value and the corresponding lump-sum fee is zero. In other words, the optimal two-part scheme with zero data cap has a pay-as-you-go structure, which is also a global optimal two-part pricing with any data cap. Since the two-part structure is a generalization of the pay-as-you-go scheme, this conclusion may not seem intuitive. We will provide a critical and detailed explanation of this result in Appendix A.

In summary, the data cap plays a transitional role between the flat-rate and pay-as-you-go pricing. Specifically, with the diminishing data cap, an ISP's optimal pricing transitions from a more flat-rate structure to a more pay-as-you-go structure, and the corresponding revenue changes from a minimum to a maximum. This impact on the transformation of optimal pricing structure holds regardless of the distribution of user demand, the level of competition and the capacity of ISP.

## V. WELFARE-OPTIMAL PRICING

In the previous section, we studied the revenue-optimal pricing under which we showed the transitional role of data cap. However, the revenue-optimal solution does not maximize the social welfare, i.e., the total utility of the providers and their users. Although market competition in general improves the social welfare, which will be maximized under a perfect competitive market, a large deviation from the maximum welfare would more likely happen in a monopoly market.

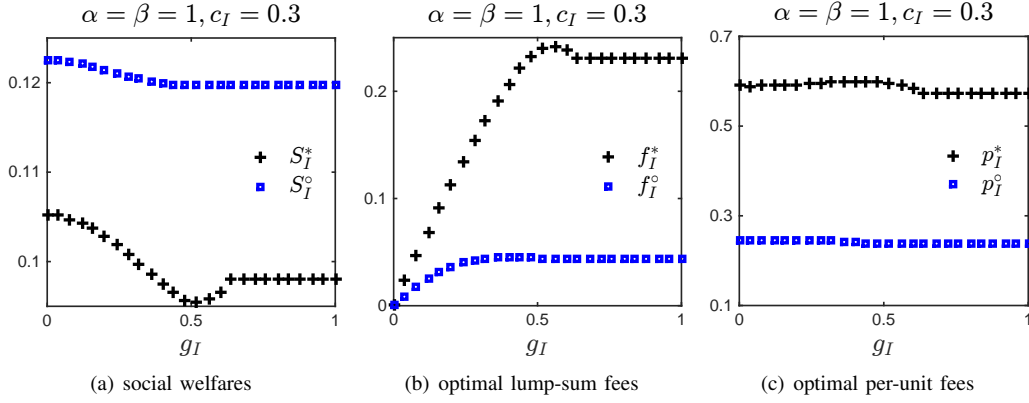


Fig. 3. Welfare-optimal and revenue-optimal pricing schemes under varying data cap.

In the section, we focus on a monopoly market, i.e., the congestion level of the free provider  $q_0 = +\infty$ . We compare the welfare-optimal and revenue-optimal solutions of the monopoly provider, and characterize the impact of data cap under the welfare-optimal pricing structure.

Under any given data cap  $g_I$ , we denote  $f_I^o(g_I)$  and  $p_I^o(g_I)$  as the optimal lump-sum fee and per-unit fee that maximize the social welfare  $S_I$  (defined in Equation (5)) and result in the optimal welfare  $S_I^o(g_I)$ . To make a comparison, we denote  $S_I^*(g_I)$  as the social welfare achieved under the provider's revenue-optimal fees  $f_I^*(g_I)$  and  $p_I^*(g_I)$ . Next, we study how various data caps influence its optimal welfare  $S_I^o$ , and corresponding pricing decisions, i.e.,  $f_I^o$  and  $p_I^o$ . In particular, we compare the two-part pricing with the pay-as-you-go scheme, i.e.,  $f_I = g_I = 0$ , and the flat-rate scheme, i.e.,  $g_I = +\infty$ . For simplification, we define the optimal social welfare under the flat-rate as  $S_\infty^o = \lim_{g_I \rightarrow \infty} S_I^o(g_I)$ .

Figure 3 shows how data cap  $g_I$  influences the optimal social welfare  $S_I^o(g_I)$ , and the welfare-optimal fees  $f_I^o, p_I^o$ , under the uniform user distribution, i.e.,  $\alpha = \beta = 1$ , and the ISP capacity  $c_I = 0.3$ . Subfigures (a) to (c) plot  $S_I^o, f_I^o$  and  $p_I^o$  as functions of the data cap  $g_I$  varying along the x-axis, respectively. As a comparison, we also plot the social welfare  $S_I^*$  and the corresponding revenue-optimal fees  $f_I^*, p_I^*$  in the three subfigures, respectively. We observe that the optimal welfare  $S_I^o$  increases monotonically with the decrease of the data cap, which has the same changing trend as the optimal revenue  $R_I^*$  shown in Figure 2. In other word, when the pricing move from a more flat-rate structure to a more usage-based structure with the diminishing data cap, the optimal welfare changes from a minimum to a maximum. Further, as the data cap  $g_I$  is zero, the lump-sum fee  $f_I^o$  is also zero, implying the welfare-optimal pricing becomes a pay-as-you-go scheme. The following theorem shows that these observations on the optimal welfare and fees are not particular to the model parameters, i.e.,  $\alpha = \beta = 1, c_I = 0.3$ .

**Theorem 4:** For any parameters  $\alpha, \beta, c_I > 0$  under Assumption 4, the optimal social welfare satisfies

$$S_\infty^o \leq S_I^o(g_I) \leq S_I^o(0) \quad \text{for } \forall g_I \geq 0.$$

Besides, when the data cap is zero, the welfare-optimal lump-sum fee is zero, i.e.,  $f_I^o(0) = 0$ .

Theorem 4 states that the social welfare generated from any optimal two-part pricing under a fixed data cap  $g_I \geq 0$  is no less than that under an optimal flat-rate scheme and no more than that under an optimal pay-as-you-go scheme. The explanation is similar to that of the result for the optimal revenue in Theorem 3, i.e., the pay-as-you-go structure is a pure usage-based pricing mode which is more fine-grained than the flat-rate structure. It can utilize the provider's capacity more effectively and generate higher welfare from the market including users of various demands. Thus, with the diminishing data cap, the welfare-optimal pricing transforms from a more flat-rate structure to a more pay-as-you-go structure, and the corresponding welfare varies from a minimum to a maximum. The results of Theorem 3 and 4 imply that both of the revenue and welfare objectives motivate the data cap decrease and move the two-part pricing towards usage-based schemes. This implication provides justifications for regulators, e.g., the US FCC, to encourage the shift to a more usage-based pricing with limited data cap for the Internet service providers.

From Figure 3, we also observe that the curves of the welfare-optimal fees  $f_I^o$  and  $p_I^o$  are always lower than those of the revenue-optimal fees  $f_I^*$  and  $p_I^*$ , respectively. Because exorbitant fees reduce users' utilities and data demand, resulting lower social welfare. This implies that regulators might need to regulate the lump-sum and per-unit fees to guarantee higher social welfare.

In summary, our results suggest that from a welfare perspective, the regulators should encourage providers to shift towards two-part pricing with limited data caps; however, the lump-sum and per-unit fees might need to be regulated when market competition does not exist.

## VI. CONCLUSIONS

In this paper, we study the role of data cap in the optimal structure of two-part tariffs. We present a novel model of users' demand and preferences over pricing and congestion alternatives and derive the market share and congestion of the



service providers under a market equilibrium. Based on the equilibrium model, we characterize the two-part structure of the revenue-optimal and welfare-optimal pricing schemes. We identify that the data cap provides a mechanism for Internet service providers to transition from the flat-rate to pay-as-you-go type of usage-based schemes. Our results reveal that both of the revenue and welfare objectives move the two-part pricing towards usage-based schemes with diminishing data caps.

#### APPENDIX A EXPLANATION OF THEOREM 3

In this section, we provide a detailed explanation of why the pay-as-you-go pricing is a global revenue-optimal two-part pricing structure. As shown in Equation (2) and (3), the selection and usage of users on the ISP depend strongly on their desirable demand. We first consider the pricing strategies on the set of users of the same demand.

Under any two-part pricing  $\theta_I = (g_I, f_I, p_I)$  and congestion level  $q_I$ , we denote  $\Phi_I(\theta_I, q_I, u)$  as the market share of the ISP on the set of users who have desirable demand  $u$ , and we denote  $R_I(\theta_I, q_I, u)$  and  $d_I(\theta_I, q_I, u)$  as the revenue and data load generated by these users, respectively, defined as

$$R_I(\theta_I, q_I, u) \triangleq \int_{\Phi_I(\theta_I, q_I, u)} t(y_I^*(\phi), \theta_I) d\mu \quad \text{and} \\ d_I(\theta_I, q_I, u) \triangleq \int_{\Phi_I(\theta_I, q_I, u)} y_I^*(\phi) d\mu$$

where  $y_I^*(\phi) = y^*(\phi, \theta_I, q_I)$  and  $t(y_I^*(\phi), \theta_I)$  are user's actual data usage and charge, respectively. To distinguish the pay-as-you-go pricing from the two-part pricing, we denote it as  $\tilde{\theta}_I \triangleq (0, 0, \tilde{p}_I)$ .

**Proposition 2:** Under Assumption 4 and any fixed congestion levels  $q_I, q_0$  and user demand  $u$ , for any two-part pricing  $\theta_I$ , there always exists a pay-as-you-go pricing  $\tilde{\theta}_I$  satisfying that  $R_I(\tilde{\theta}_I, q_I, u) = R_I(\theta_I, q_I, u)$  and  $d_I(\tilde{\theta}_I, q_I, u) \leq d_I(\theta_I, q_I, u)$ .

Proposition 2 states that under a fixed congestion level, for any two-part pricing scheme, an ISP could always find a pay-as-you-go scheme which generates the same revenue but lower data load from the set of users of the same desirable demand. It implies that the pay-as-you-go pricing is the most efficient two-part pricing structure for the set of users of the same demand. To further analyze the pricing schemes in the market including users of various demands, we define a *demand-based* two-part pricing, denoted as  $\theta_I(u) \triangleq (g_I(u), f_I(u), p_I(u))$  ( $u \in [0, U]$ ), where  $g_I(u), f_I(u)$  and  $p_I(u)$  are continuous functions of the desirable demand  $u$ . Under this pricing strategy, a user with desirable demand  $u_0 \in [0, U]$  is charged based on the data cap  $g_I(u_0)$ , the lump-sum fee  $f_I(u_0)$  and the per-unit fee  $p_I(u_0)$ . This demand-based pricing structure is a generalization of the two-part pricing  $\theta_I = (g_I, f_I, p_I)$ , i.e.,  $g_I(u) = g_I, f_I(u) = f_I, p_I(u) = p_I$  for any  $u \in [0, U]$ . To distinguish the two types of pricing modes, we call  $\theta_I$  as a *fixed* two-part pricing. Similarly, we refer to  $\tilde{\theta}_I(u) \triangleq (0, 0, \tilde{p}_I(u))$  and  $\tilde{\theta}_I = (0, 0, \tilde{p}_I)$  as *demand-based* and *fixed* pay-as-you-go pricing, respectively, where the former is a generalization of the latter.

Based on Definition 2, we extend the definition of congestion equilibrium from the fixed two-part pricing to the demand-based two-part pricing, as follows.

**Definition 3:** Given the congestion level  $q_0$ , for the ISP  $I$  with any fixed demand-based two-part pricing  $\theta_I(u)$  and capacity  $c_I$ ,  $q_I$  is an equilibrium if and only if

$$q_I = Q_I\left(D(\Phi_I(\theta_I(u), q_I, q_0); \theta_I(u), q_I), c_I\right)$$

where  $\Phi_I$  is the set of users who choose the ISP and the function  $D$  is the aggregate data usage (load) of the users.

**Lemma 1:** Under Assumption 1-3, for any given demand-based pricing strategy  $\theta_I(u)$ , capacity  $c_I$  and congestion  $q_0$ , there always exists a unique equilibrium  $q_I$ , satisfying  $q_I < q_0$ .

Lemma 1 guarantees the existence and uniqueness of the congestion equilibrium under any demand-based pricing strategy. We denote the equilibrium under the pricing strategy  $\theta_I(u)$  as  $q_I(\theta_I(u))$ . Further, we denote the provider's revenue and data load from all users of various demands in equilibrium as  $R_I(\theta_I(u), q_I(\theta_I(u)))$  and  $d_I(\theta_I(u), q_I(\theta_I(u)))$ , respectively, defined by

$$R_I(\theta_I(u), q_I(\theta_I(u))) = \int_0^U R_I(\theta_I(u), q_I(\theta_I(u)), u) du \quad \text{and} \\ d_I(\theta_I(u), q_I(\theta_I(u))) = \int_0^U d_I(\theta_I(u), q_I(\theta_I(u)), u) du$$

where  $R_I(\theta_I(u), q_I(\theta_I(u)), u)$  and  $d_I(\theta_I(u), q_I(\theta_I(u)), u)$  are the revenue and data load generated from the users of desirable demand  $u$ .

**Proposition 3:** Under Assumption 4, for any demand-based two-part pricing  $\theta_I(u)$ , there always exists a demand-based pay-as-you-go pricing  $\tilde{\theta}_I(u)$  satisfying that  $R_I(\tilde{\theta}_I(u), q_I(\tilde{\theta}_I(u))) \geq R_I(\theta_I(u), q_I(\theta_I(u)))$ .

Proposition 3 tells that under equilibrium, for any demand-based two-part pricing scheme, there must exist a demand based pay-as-you-go pricing which could generate a higher or the same revenue from all users of various demands. The reason is that from Proposition 2, for the given demand-based two-part pricing  $\theta_I(u)$ , we could always find a demand-based pay-as-you-go pricing  $\tilde{\theta}_I(u)$  which generates the same revenue but no larger data load from users of various demands, under the same congestion level  $q_I(\theta_I(u))$ . Because the congestion level decreases with the data load based on Assumption 3, the congestion equilibrium  $q_I(\tilde{\theta}_I(u))$  must be no larger than the congestion  $q_I(\theta_I(u))$ . Further, because the provider's revenue is non-increasing in its congestion level, the ISP's revenue under the pay-as-you-go pricing  $\tilde{\theta}_I(u)$  is no less than that under the two-part pricing  $\theta_I(u)$  in equilibrium.

Proposition 3 implies that the demand-based pay-as-you-go pricing is an optimal demand-based two-part pricing structure. However, it is not enough to declare that the fixed pay-as-you-go and two-part pricing also satisfy this relations. To achieve this, we provide the following proposition.

**Proposition 4:** Under Assumption 4, if  $\tilde{\theta}_I \triangleq (0, 0, \tilde{p}_I)$  is a revenue-optimal fixed pay-as-you-go pricing, then  $\tilde{\theta}_I(u) \triangleq$



$(0, 0, \tilde{p}_I(u))$  where  $\tilde{p}_I(u) = \tilde{p}_I$  for  $\forall u \in [0, U]$  must be a revenue-optimal demand-based pay-as-you-go pricing.

Proposition 4 shows that an optimal fixed pay-as-you-go pricing scheme is also an optimal demand-based pay-as-you-go pricing scheme. Intuitively, the pay-as-you-go structure is a pure usage-based pricing mode, i.e., the charge for a user is proportional to her usage. Thus, the revenue-optimal price only depends on the distribution of user value. Because the user's value and demand are independent from Assumption 4, under the demand-based pay-as-you-go pricing structure, the revenue-optimal prices for the user sets of different desirable demands are all the same, which is also a revenue-optimal price under the fixed pay-as-you-go pricing structure.

So far, we can provide a full explanation of why the fixed pay-as-you-go pricing is a revenue-optimal fixed two-part pricing structure. Based on Proposition 4, an optimal fixed pay-as-you-go scheme is also an optimal demand-based pay-as-you-go scheme, which generates no lower revenue compared with any demand-based two-part scheme from Proposition 3. Because the fixed two-part pricing is a specialization of the demand-based pricing, the optimal fixed pay-as-you-go scheme obtains no lower revenue than any fixed two-part scheme, i.e., the pay-as-you-go pricing is a revenue-optimal two-part pricing structure.

## APPENDIX B PROOFS OF ANALYTICAL RESULTS

**Proof of Proposition 1:** For a set  $\mathcal{N}$  of providers and any user type  $\phi \in \Phi_i(\theta, \mathbf{q}, q_0)$ , we have  $y^*(\phi, \hat{\theta}_i, q_i) \geq y^*(\phi, \theta_i, q_i)$ . Because the optimal data usage  $y^*$  is non-increasing in  $p$  and non-decreasing in  $g$  from Equation (2). Further, the utility function  $\pi(y)$  is non-decreasing in the data usage  $y$ , thus it satisfies  $\pi(y^*(\phi, \hat{\theta}_i, q_i)) \geq \pi(y^*(\phi, \theta_i, q_i)) \geq \pi(y^*(\phi, \theta_j, q_j)) = \pi(y^*(\phi, \hat{\theta}_j, q_j))$  for  $\forall j \neq i$ . Then we have  $\phi \in \Phi_i(\hat{\theta}, \mathbf{q}, q_0)$  from Definition 1 and therefore  $\Phi_i(\theta, \mathbf{q}, q_0) \subseteq \Phi_i(\hat{\theta}, \mathbf{q}, q_0)$ . Similarly, we can show that  $\Phi_j(\theta, \mathbf{q}, q_0) \supseteq \Phi_j(\hat{\theta}, \mathbf{q}, q_0), \forall j \neq i$ .

For two sets  $\mathcal{N}$  and  $\mathcal{N}'$  of providers and any provider  $i \in \mathcal{N}$ , for any user type  $\phi \in \Phi_i(\theta', \mathbf{q}', q_0)$ , based on Definition 1, it satisfies that  $\pi(y^*(\phi, \theta_i, q_i)) = \pi(y^*(\phi, \theta'_i, q'_i)) \geq \pi(y^*(\phi, \theta'_j, q'_j)) = \pi(y^*(\phi, \theta_j, q_j)), \forall j \in \mathcal{N} \setminus \{i\}$ . Then we have  $\phi \in \Phi_i(\theta, \mathbf{q}, q_0)$  and therefore  $\Phi_i(\theta', \mathbf{q}', q_0) \subseteq \Phi_i(\theta, \mathbf{q}, q_0), \forall i \in \mathcal{N}$ . ■

**Proof of Theorem 1:** We first prove the existence of equilibrium. By Definition 2,  $\mathbf{q}$  is an equilibrium if and only if for all  $i \in \mathcal{N}$ ,

$$q_i = Q_i(D(\Phi_i(\theta, \mathbf{q}, q_0); \theta_i, q_i), c_i) = Q_i(D(\theta, \mathbf{q}, q_0), c_i).$$

Since  $\theta, \mathbf{c}$  and  $q_0$  are constants, we omit them in the notation and write the above in a matrix form as  $\mathbf{q} = Q(D(\mathbf{q})) = Q \circ D(\mathbf{q})$ . Thus, we can view the composite function  $Q \circ D$  as a mapping from the convex set  $\mathbb{R}_+^{|\mathcal{N}|}$  to itself. By Assumption 3, we know that each  $Q_i(d_i, c_i)$  is continuous in  $d_i$  and thus each  $Q_i(D(\theta, \mathbf{q}, q_0), c_i)$  is continuous in  $\mathbf{q}$ . To this end, we know that  $Q \circ D(\mathbf{q})$  is continuous in  $\mathbf{q}$ . By Brouwer fixed-

point theorem, there always exists a fixed point that satisfies  $Q \circ D(\mathbf{q}) = \mathbf{q}$  and is also an equilibrium.

We then prove  $q_i < q_0$  ( $\forall i \in \mathcal{N}$ ) under equilibrium by contradiction. Suppose there exists a fixed pricing decision  $\theta$ , capacity  $\mathbf{c}$  and congestion  $q_0$  that make  $q_i \geq q_0 > 0$  under equilibrium. For any user type  $\phi$ , we have  $\pi(y_i^*(\phi)) = vy_i^*(\phi) - t(y_i^*(\phi), \theta_i) \leq vy_i^*(\phi) \leq v\rho(u, q_i) \leq v\rho(u, q_0) = \pi(y_0^*(\phi))$ , i.e., the free provider induces higher maximum utility than the provider  $i$ . Thus no user would choose provider  $i$  and its congestion  $q_i = Q_i(0, c_i) = 0$  from Assumption 3, which is contradictory with the supposition that  $q_i \geq q_0 > 0$ .

Finally, we prove the uniqueness of the equilibrium when the market has only one normal provider  $I$  by contradiction. Suppose there exist two equilibriums  $q_I$  and  $q'_I$  under fixed pricing strategy  $\theta_I$ , capacity  $c_I$  and congestion  $q_0$ . Without loss of generality, we assume that  $q'_I > q_I$ . For any user type  $\phi \in \Phi_I(\theta_I, q'_I, q_0)$ , we have  $y^*(\phi, \theta_I, q_I) \geq y^*(\phi, \theta_I, q'_I)$ . Because the optimal data usage  $y^*$  is non-increasing in the congestion  $q_I$ . Further, the utility function  $\pi(y)$  is non-decreasing in the data usage  $y$ , thus it satisfies  $\pi(y^*(\phi, \theta_I, q_I)) \geq \pi(y^*(\phi, \theta_I, q'_I)) \geq \pi(y^*(\phi, \theta_0, q_0))$  implying that  $\phi \in \Phi_I(\theta_I, q_I, q_0)$  from Definition 1. Thus we have  $\Phi_I(\theta_I, q'_I, q_0) \subseteq \Phi_I(\theta_I, q_I, q_0)$ . By Equation (4), we deduce  $D(\theta_I, q_I, q_0) = \int_{\Phi_I(\theta_I, q_I)} y^*(\phi, \theta_I, q_I) d\mu \geq \int_{\Phi_I(\theta_I, q'_I)} y^*(\phi, \theta_I, q'_I) d\mu = D(\theta_I, q'_I, q_0)$ . As  $Q_I(d_I, c_I)$  is non-decreasing in  $d_I$  from Assumption 3, we have

$$q_I = Q_I(D(\theta_I, q_I, q_0), c_I) \geq Q_I(D(\theta_I, q'_I, q_0), c_I) = q'_I$$

based on Definition 2, which is contradictory with the supposition that  $q'_I > q_I$ . Therefore, the equilibrium is unique in the market with only one normal provider. ■

**Proof of Theorem 2:** We first prove the congestion  $q_I(\theta_I, c_I, q_0)$  is non-increasing in the lump-sum fee  $f_I$  by contradiction. Suppose  $q_I(\theta_I, c_I, q_0)$  is not non-increasing in  $f_I$ , there must exist pricing strategies  $\theta'_I = (g_I, f'_I, p_I)$  and  $\theta_I = (g_I, f_I, p_I)$  satisfying  $f'_I > f_I$  and  $q_I(\theta'_I, c_I, q_0) > q_I(\theta_I, c_I, q_0)$ . For any user type  $\phi \in \Phi_I(\theta'_I, q_I(\theta'_I, c_I, q_0), q_0)$ , because the optimal data usage  $y^*$  is non-increasing in the congestion  $q$ , we have  $y^*(\phi, \theta'_I, q_I(\theta'_I, c_I, q_0)) \leq y^*(\phi, \theta_I, q_I(\theta_I, c_I, q_0))$ . Furthermore, because the utility function  $\pi$  is non-decreasing in  $y$  and non-increasing in  $f$ , we have

$$\pi(y^*(\phi, \theta_I, q_I(\theta_I, c_I, q_0))) \geq \pi(y^*(\phi, \theta'_I, q_I(\theta'_I, c_I, q_0)))$$

implying that  $\phi \in \Phi_I(\theta_I, q_I(\theta_I, c_I, q_0), q_0)$  from Definition 1. Thus  $\Phi_I(\theta'_I, q_I(\theta'_I, c_I, q_0), q_0) \subseteq \Phi_I(\theta_I, q_I(\theta_I, c_I, q_0), q_0)$ . From Equation (4), it satisfies  $D(\theta'_I, q_I(\theta'_I, c_I, q_0), q_0) \leq D(\theta_I, q_I(\theta_I, c_I, q_0), q_0)$ . Because  $Q_I(d_I, c_I)$  is non-decreasing in  $d_I$  from Assumption 3, we have

$$\begin{aligned} q_I(\theta'_I, c_I, q_0) &= Q_I(D(\theta'_I, q_I(\theta'_I, c_I, q_0), q_0), c_I) \\ &\leq Q_I(D(\theta_I, q_I(\theta_I, c_I, q_0), q_0), c_I) = q_I(\theta_I, c_I, q_0) \end{aligned}$$

by Definition 2, which is contradictory with the supposition that  $q_I(\theta'_I, c_I, q_0) > q_I(\theta_I, c_I, q_0)$ . Therefore,  $q_I(\theta_I, c_I, q_0)$

is non-increasing in  $f_I$ . Similarly, we can prove that  $q_I(\theta_I, c_I, q_0)$  is non-increasing in the per-unit fee  $p_I$ , the capacity  $c_I$  and non-decreasing in the data cap  $g_I$ , the congestion level  $q_0$  and when new providers enter the market, the existing provider's congestion must be improved, i.e.,  $q_I(\theta, c, q_0) \leq q_I(\theta_I, c_I, q_0)$ . ■

**Proof of Corollary 1:** Given  $\hat{c} = kc/U$ ,  $Q_i(d_i, c_i) = d_i/c_i$  and  $\hat{d}_i = kd_i/U$  for all  $i \in \mathcal{N}$ , we know that  $\hat{q}_i = Q(\hat{d}_i, \hat{c}_i) = \hat{d}_i/\hat{c}_i = (kd_i/U)/(kc_i/U) = Q_i(d_i, c_i) = q_i$ . Thus, we only need to show that under  $\hat{\mathbf{q}} = \mathbf{q}$ , we must have  $\hat{d}_i = kd_i/U$  for all  $i \in \mathcal{N}$ . We denote  $\hat{\Phi}_i$  as the market share of the provider  $i$  in the normalized system. For simplification, we denote  $\hat{\theta}_i = (\hat{g}_i, \hat{f}_i, \hat{p}_i)$ ,  $\hat{\theta}_0 = (0, 0, 0)$ . For any  $\phi = (u, v) \in \Phi$ , we denote  $\hat{\phi} = (\hat{u}, \hat{v}) = (u/U, v/V) \in \hat{\Phi}$ . Then it satisfies  $y^*(\phi, \theta_i, q_i) = \rho(u, q_i) - [\rho(u, q_i) - g_i]^+ \mathbf{1}_{\{v < p_i\}} = ue^{-q_i} - (ue^{-q_i} - g_i)^+ \mathbf{1}_{\{v < p_i\}} = U[ue^{-q_i}/U - (ue^{-q_i}/U - g_i/U)^+ \mathbf{1}_{\{v/V < p_i/V\}}] = U[\hat{u}e^{-\hat{q}_i} - (\hat{u}e^{-\hat{q}_i} - \hat{g}_i)^+ \mathbf{1}_{\{\hat{v} < \hat{p}_i\}}] = Uy^*(\hat{\phi}, \hat{\theta}_i, \hat{q}_i)$ . Because it satisfies  $\pi(y^*(\hat{\phi}, \hat{\theta}_i, \hat{q}_i)) - \pi(y^*(\hat{\phi}, \hat{\theta}_j, \hat{q}_j)) = \hat{v}[y^*(\hat{\phi}, \hat{\theta}_i, \hat{q}_i) - y^*(\hat{\phi}, \hat{\theta}_j, \hat{q}_j)] - t(y^*(\hat{\phi}, \hat{\theta}_i, \hat{q}_i), \hat{\theta}_i) + t(y^*(\hat{\phi}, \hat{\theta}_j, \hat{q}_j), \hat{\theta}_j) = v[y^*(\phi, \theta_i, q_i) - y^*(\phi, \theta_j, q_j)]/UV - t(y^*(\phi, \theta_i, q_i), \theta_i)/UV + t(y^*(\phi, \theta_j, q_j), \theta_j)/UV = \pi(y^*(\phi, \theta_i, q_i))/UV - \pi(y^*(\phi, \theta_j, q_j))/UV$ , we have  $\hat{\phi} \in \hat{\Phi}_i(\hat{\theta}, \hat{\mathbf{q}}, \hat{q}_0)$  if and only if  $\phi \in \Phi_i(\theta, \mathbf{q}, q_0)$ . Further, we can deduce that  $\hat{d}_i = \int_{\hat{\Phi}_i} y^*(\hat{\phi}, \hat{\theta}_i, \hat{q}_i) d\hat{\mu} = \int_{\Phi_i} y^*(\phi, \theta_i, q_i)/U dk\mu = kd_i/U$ . Therefore, Corollary 1 concludes. ■

**Proof of Theorem 3:** We first prove that  $R_I^*(g_I) \geq R_\infty^*$  for any  $g_I \geq 0$ . Because the maximum desirable demand  $U$  of all users is normalized to 1, we have  $R_\infty^* = R_I^*(1)$  where  $R_I^*(1)$  is the ISP's optimal revenue under the flat-rate pricing. We denote  $f_I^*(1)$  as any optimal lump-sum fee that maximizes the revenue under the flat-rate. We denote  $R_I(g_I, f_I^*(1), 0)$  as the ISP's revenue under the pricing decision  $\theta_I = (g_I, f_I^*(1), 0)$  for any data cap  $g_I \in [0, 1]$ . Because It satisfies  $R_I^*(g_I) \geq R_I(g_I, f_I^*(1), 0) = R_I^*(1)$ , we have  $R_I^*(g_I) \geq R_\infty^*$  for any  $g_I$ .

We then prove that  $R_I^*(g_I) \leq R_I^*(0)$ . We only need to show that any revenue-optimal (fixed) pay-as-you-go pricing is also a global revenue-optimal (fixed) two-part pricing. This result can be implied Based on Proposition 3 and 4 whose proofs could be found later. From Proposition 4, a revenue-optimal fixed pay-as-you-go pricing is also a revenue-optimal demand-based pay-as-you-go pricing, which generates higher or the same revenue compared with any demand-based two-part pricing from Proposition 3. Because the fixed two-part pricing structure is a specialization of the demand-based one, any revenue-optimal fixed pay-as-you-go pricing obtains higher or the same revenue than any fixed two-part pricing, i.e.,  $R_I^*(g_I) \leq R_I^*(0)$ .

We finally prove that  $f_I^*(0) = 0$ . We only needs to show that any pricing scheme  $\theta_I = (0, f_I, p_I)$  ( $f_I > 0$ ) must not be a global revenue-optimal two-part pricing. For any two-part pricing scheme  $\theta_I = (0, f_I, p_I)$  ( $f_I > 0$ ), the pay-as-you-go pricing  $\tilde{\theta}_I(u) = (0, 0, f_I/\rho(u, q_I(\theta_I)) + p_I)$  satisfies that  $R_I(\tilde{\theta}_I(u), q_I(\tilde{\theta}_I(u))) = R_I(\theta_I, q_I(\theta_I))$ . From the proof of Proposition 4, a demand-based pay-as-you-go pricing  $\tilde{\theta}_I(u) =$

$(0, 0, \tilde{p}_I(u))$  is revenue-optimal only if the per-unit fee  $\tilde{p}_I(u)$  is a constant function of  $u$ . Thus,  $(0, 0, f_I + p_I\rho(u, q_I(\theta_I)))$  must not be a revenue-optimal demand-based pay-as-you-go pricing, and further, the two-part pricing  $(0, f_I, p_I)$  ( $f_I > 0$ ) must not be a global revenue-optimal two-part pricing.

In summary, Theorem 3 concludes. ■

**Proof of Proposition 2:** For any fixed congestion levels  $q_I, q_0$ , user demand  $u$ , and any two-part pricing  $\theta_I$ , we conduct the pay-as-you-go pricing  $\tilde{\theta}_I = (0, 0, \tilde{p}_I)$  which satisfies that  $R_I(\tilde{\theta}_I, q_I, u) = R_I(\theta_I, q_I, u)$  and  $d_I(\tilde{\theta}_I, q_I, u) \leq d_I(\theta_I, q_I, u)$  by two cases as follows.

The first case:  $\rho(u, q_I) \leq g_I$  or  $g_I p_I - f_I \leq \rho(u, q_0) \cdot p_I$  or  $q_I \geq q_0$ . Under this case, we set

$$\tilde{p}_I = \frac{f_I + [\rho(u, q_I) - g_I]^+ p_I}{\rho(u, q_I)} \quad (7)$$

Then based on Definition 1 and Assumption 2, when any user of desirable demand  $u$  faces the pricing strategies  $\theta_I$  and  $\tilde{\theta}_I$ , respectively, if she chooses to use the ISP, her actual usages are both  $\rho(u, q_I)$  and the charges are  $f_I + [\rho(u, q_I) - g_I]^+ p_I$  and  $\rho(u, q_I)\tilde{p}_I$ , which are the same from Equation (7). Thus, the ISP could obtain the same market share and further the same revenue and data load from the set of users of desirable demand  $u$  when it adopts the pricing strategies  $\theta_I$  and  $\tilde{\theta}_I$ , respectively, i.e.,  $R_I(\theta_I, q_I, u) = R_I(\tilde{\theta}_I, q_I, u)$  and  $d_I(\theta_I, q_I, u) = d_I(\tilde{\theta}_I, q_I, u)$ .

The second case:  $\rho(u, q_I) > g_I$  and  $g_I p_I - f_I > \rho(u, q_0) \cdot p_I$  and  $q_I < q_0$ . Under this case, when the ISP adopts the two-part pricing scheme  $\theta_I$ , its revenue from the set of users of desirable demand  $u$  based on Equation (5) is

$$R_I(\theta_I, q_I, u) = \left[ (g_I - \rho(u, q_0)) \int_{\frac{f_I}{g_I - \rho(u, q_0)}}^1 \frac{f_I}{g_I - \rho(u, q_0)} dF_v + (\rho(u, q_I) - g_I) \int_{p_I}^1 p_I dF_v \right] F'_u(u). \quad (8)$$

Based on Assumption 4, the distribution function of user valuation is  $F_v(v) = v^\beta$ . We define a continuous function  $y(x) \triangleq \int_x^1 x dF_v(x) = x(1 - x^\beta)$ . It is decreasing in  $x \in [(\frac{1}{\beta+1})^{\frac{1}{\beta}}, 1]$  and satisfies that

$$0 = y(1) \leq y(x) \leq y\left(\left(\frac{1}{\beta+1}\right)^{\frac{1}{\beta}}\right) \quad \text{for } \forall x \in [0, 1].$$

Combining with Equation (8), we have

$$R_I(\theta_I, q_I, u) \leq [\rho(u, q_I) - \rho(u, q_0)] y\left(\left(\frac{1}{\beta+1}\right)^{\frac{1}{\beta}}\right) F'_u(u). \quad (9)$$

We set the solution of the following Equation (10) as the pay-as-you-go pricing scheme  $\tilde{\theta}_I$ ,

$$\begin{cases} R_I(\tilde{\theta}_I, q_I, u) = R_I(\theta_I, q_I, u), \\ \left(\frac{1}{\beta+1}\right)^{\frac{1}{\beta}} \leq M \cdot \tilde{p}_I \leq 1 \end{cases} \quad (10)$$

where  $M = \frac{\rho(u, q_I)}{\rho(u, q_I) - \rho(u, q_0)}$  and the ISP's revenue  $R_I(\tilde{p}_I, q_I, u)$  from the set of users of desirable demand  $u$

based on Equation (5) is

$$R_I(\tilde{p}_I, q_I, u) = [\rho(u, q_I) - \rho(u, q_0)]y(M\tilde{p}_I)F'_u(u). \quad (11)$$

Combining Equation (9) and (11), we know Equation (10) must exist a unique solution, because  $y(x)$  is decreasing in  $x \in [(\frac{1}{\beta+1})^{\frac{1}{\beta}}, 1]$  and  $y(1) = 0$ .

Next, we show that  $d_I(\theta_I, q_I, u) \geq d_I(\tilde{\theta}_I, q_I, u)$  under two situations: (i)  $p_I \leq M\tilde{p}_I$  and (ii)  $p_I > M\tilde{p}_I$ .

For Situation (i), it satisfies  $\frac{f_I}{g_I - \rho(u, q_0)} < p_I \leq M\tilde{p}_I$ . From Equation (4), we have

$$\begin{aligned} d_I(\tilde{\theta}_I, q_I, u)/F'_u(u) &= [\rho(u, q_I) - \rho(u, q_0)] \int_{M\tilde{p}_I}^1 dF_v \\ &\leq [\rho(u, q_I) - g_I] \int_{p_I}^1 dF_v + [g_I - \rho(u, q_0)] \int_{\frac{f_I}{g_I - \rho(u, q_0)}}^1 dF_v \\ &= d_I(\theta_I, q_I, u)/F'_u(u). \end{aligned}$$

For Situation (ii), we first define a continuous function  $z(x) = \int_x^1 dF_v = 1 - x^\beta$ , then we have  $y(z(x)) = \int_x^1 x dF_v = z(x)(1 - z(x))^{\frac{1}{\beta}}$ . Further,  $y(z)$  is increasing and concave in  $z \in [0, \frac{\beta}{\beta+1}]$  and decreasing in  $z \in (\frac{\beta}{\beta+1}, 1)$ . If  $\frac{f_I}{g_I - \rho(u, q_0)} \geq (\frac{1}{\beta+1})^{\frac{1}{\beta}}$ , we have  $0 \leq z(p_I), z(M\tilde{p}_I), z(\frac{f_I}{g_I - \rho(u, q_0)}) \leq \frac{\beta}{\beta+1}$ . From Equation (8), (10) and (11), we know that

$$\begin{aligned} &[\rho(u, q_I) - \rho(u, q_0)]y(z(M\tilde{p}_I)) = [\rho(u, q_I) - g_I]y(z(p_I)) \\ &+ [g_I - \rho(u, q_0)]y(z(\frac{f_I}{g_I - \rho(u, q_0)})). \end{aligned}$$

Because  $y(z)$  is increasing and concave in  $z \in [0, \frac{\beta}{\beta+1}]$ , it satisfies that

$$\begin{aligned} d_I(\tilde{\theta}_I, q_I, u)/F'_u(u) &= [\rho(u, q_I) - \rho(u, q_0)]z(M\tilde{p}_I) \\ &\leq [\rho(u, q_I) - g_I]z(p_I) + [g_I - \rho(u, q_0)]z(\frac{f_I}{g_I - \rho(u, q_0)}) \\ &= d_I(\theta_I, q_I, u)/F'_u(u). \end{aligned}$$

implying that  $d_I(\tilde{\theta}_I, q_I, u) \leq d_I(\theta_I, q_I, u)$ . If  $\frac{f_I}{g_I - \rho(u, q_0)} < (\frac{1}{\beta+1})^{\frac{1}{\beta}}$ , there must exist a value  $k_I \in [(\frac{1}{\beta+1})^{\frac{1}{\beta}}, 1]$  which satisfies that  $y(z(k_I)) = y(z(\frac{f_I}{g_I - \rho(u, q_0)}))$ . Then we have  $0 \leq z(p_I), z(M\tilde{p}_I), z(k_I) \leq \frac{\beta}{\beta+1}$  and  $z(k_I) \leq z(\frac{f_I}{g_I - \rho(u, q_0)})$ . From Equation (8), (10) and (11), we have that

$$\begin{aligned} &[\rho(u, q_I) - \rho(u, q_0)]y(z(M\tilde{p}_I)) \\ &= [g_I - \rho(u, q_0)]y(z(\frac{f_I}{g_I - \rho(u, q_0)})) + [\rho(u, q_I) - g_I]y(z(p_I)) \\ &= [g_I - \rho(u, q_0)]y(z(k_I)) + [\rho(u, q_I) - g_I]y(z(p_I)). \end{aligned}$$

Because  $y(z)$  is increasing and concave in  $z \in [0, \frac{\beta}{\beta+1}]$ , it satisfies that

$$\begin{aligned} d_I(\tilde{\theta}_I, q_I, u)/F'_u(u) &= [\rho(u, q_I) - \rho(u, q_0)]y(z(M\tilde{p}_I)) \\ &\leq [g_I - \rho(u, q_0)]z(k_I) + [\rho(u, q_I) - g_I]z(p_I) \\ &\leq [g_I - \rho(u, q_0)]z(\frac{f_I}{g_I - \rho(u, q_0)}) + [\rho(u, q_I) - g_I]z(p_I) \\ &= d_I(\theta_I, q_I, u)/F'_u(u) \end{aligned}$$

implying that  $d_I(\tilde{\theta}_I, q_I, u) \leq d_I(\theta_I, q_I, u)$ .

Summarizing the two cases, Proposition 2 concludes. ■

**Proof of Lemma 1:** This lemma can be proofed by the same method with Theorem 1. In other word, for any demand-based pricing  $\theta_I(u)$ , we can show the existence of the equilibrium based on the Brouwer fixed-point theorem and the uniqueness of the equilibrium by contradiction. ■

**Proof of Proposition 3:** From Proposition 2, for any fixed  $u \in [0, 1]$  and the given demand-based two-part pricing  $\theta_I(u)$ , we could always find a demand-based pay-as-you-go pricing  $\tilde{\theta}_I(u)$  which generate the same revenue but lower data load from the set of users of various demand  $u$ , under the same congestion level  $q_I(\theta_I(u))$ , i.e., we have

$$\begin{cases} R_I(\tilde{\theta}_I(u), q_I(\theta_I(u)), u) = R_I(\theta_I(u), q_I(\theta_I(u)), u), \\ d_I(\tilde{\theta}_I(u), q_I(\theta_I(u)), u) \leq d_I(\theta_I(u), q_I(\theta_I(u)), u). \end{cases} \quad (12)$$

We make the demand-based pay-as-you-go pricing  $\tilde{\theta}_I(u)$  satisfies Equation (12) for any  $u \in [0, 1]$  and we have that

$$\begin{aligned} R_I(\tilde{\theta}_I(u), q_I(\theta_I(u))) &= \int_0^1 R_I(\tilde{\theta}_I(u), q_I(\theta_I(u)), u) du \\ &= \int_0^1 R_I(\tilde{\theta}_I(u), q_I(\theta_I(u)), u) du = R_I(\theta_I(u), q_I(\theta_I(u))), \\ d_I(\tilde{\theta}_I(u), q_I(\theta_I(u))) &= \int_0^1 d_I(\tilde{\theta}_I(u), q_I(\theta_I(u)), u) du \\ &\leq \int_0^1 d_I(\theta_I(u), q_I(\theta_I(u)), u) du = d_I(\theta_I(u), q_I(\theta_I(u))) \end{aligned} \quad (13)$$

implying that the equilibrium congestion  $q_I(\tilde{\theta}_I(u))$  under the pricing  $\tilde{\theta}_I(u)$  is no larger than the congestion  $q_I(\theta_I(u))$ . Because if we assume that  $q_I(\tilde{\theta}_I(u))$  is larger than  $q_I(\theta_I(u))$ , for the pricing scheme  $\tilde{\theta}_I(u)$ , any user's utility and usage under the congestion  $q_I(\tilde{\theta}_I(u))$  would not be larger than that under the congestion  $q_I(\theta_I(u))$ . Thus, the ISP's data load under the congestion  $q_I(\theta_I(u))$  must not be larger than that under the congestion  $q_I(\tilde{\theta}_I(u))$ , i.e.,

$$d_I(\tilde{\theta}_I(u), q_I(\tilde{\theta}_I(u))) \leq d_I(\tilde{\theta}_I(u), q_I(\theta_I(u))).$$

Combining with Equation (13), it satisfies

$$d_I(\tilde{\theta}_I(u), q_I(\tilde{\theta}_I(u))) \leq d_I(\theta_I(u), q_I(\theta_I(u))).$$

Because the level of congestion is increasing in the data load from Assumption 3, the equilibrium congestion for the pricing  $\tilde{\theta}_I(u)$  should also be no larger than that the pricing  $\theta_I(u)$ , which is contradictory with the assumption. Thus, we have  $q_I(\tilde{\theta}_I(u)) \leq q_I(\theta_I(u))$ . Further, because the provider's revenue is non-increasing in its congestion level under any given pricing strategy, we have

$$\begin{aligned} R_I(\tilde{\theta}_I(u), q_I(\tilde{\theta}_I(u))) &\geq R_I(\tilde{\theta}_I(u), q_I(\theta_I(u))) \\ &= R_I(\theta_I(u), q_I(\theta_I(u))). \end{aligned}$$

implying that the ISP could obtain no smaller revenue from the

pricing  $\tilde{\theta}_I(u)$  compared to the pricing  $\theta_I(u)$  under equilibrium. Therefore, Proposition 3 concludes. ■

**Proof of Proposition 4:** We first prove that any revenue-optimal demand-based pay-as-you-go pricing strategy  $\tilde{\theta}_I(u) = (0, 0, \tilde{p}_I(u))$  satisfies that  $\tilde{p}_I(u)$  is a constant function of  $u \in [0, 1]$ . Under Assumption 4, we denote  $h(u) \triangleq (\tilde{p}_I(u))^\beta$  for any  $u \in [0, 1]$  and it is a solution of the following maximization problem:

$$\begin{aligned} & \text{Maximize} && T(h(u)), \\ & \text{subject to} && h(u) \geq 0 \quad \text{for } \forall u \in [0, 1]. \end{aligned}$$

where the function  $T(h(u))$  is denoted as the ISP's revenue under the pay-as-you-go pricing  $\tilde{\theta}_I(u) = (0, 0, (h(u))^{\frac{1}{\beta}})$ . Thus, it satisfies that

$$T(h(u)) \geq T(h(u) + \epsilon\eta(u)) \triangleq H(\epsilon) \quad (14)$$

where  $\eta(u)$  is a disturbing function, from which we have  $\frac{dH(\epsilon)}{d\epsilon}|_{\epsilon=0} = 0$ . We denote  $s(h(u)) \triangleq q_I(\tilde{\theta}_I(u))$  as the congestion equilibrium under the pricing  $\tilde{\theta}_I(u) = (0, 0, (h(u))^{\frac{1}{\beta}})$ . From Assumption 4 and Definition 3, we have

$$\begin{aligned} c_I \cdot s(h(u)) &= \int_0^1 \int_{K(s(h(u)))}^{(h(u))^{\frac{1}{\beta}}} u dF_u \cdot e^{-s(h(u))} \\ &= \int_0^1 [1 - K^\beta(s(h(u)))h(u)] u dF_u \cdot e^{-s(h(u))} \end{aligned}$$

where  $K(x) = e^{-x}/(e^{-x} - e^{-q_0})$ . Further we derive that

$$\begin{aligned} \int_0^1 h(u) u dF_u &= \left[ \int_0^1 u dF_u - c_I s(h(u)) e^{s(h(u))} \right] \frac{1}{K^\beta(s(h(u)))} \\ &\triangleq G(s(h(u))) \end{aligned}$$

from which we have that  $s(h(u)) = G^{-1}(\int_0^1 h(u) u dF_u)$  where  $G^{-1}$  is the inverse function of  $G$ .

For any function  $\eta(u)$  with condition  $\int_0^1 u \eta(u) dF_u = 0$ , we have

$$\begin{aligned} & \frac{ds(h(u) + \epsilon\eta(u))}{d\epsilon} \Big|_{\epsilon=0} \\ &= (G^{-1})' \left( \int_0^1 h(u) u dF_u \right) \cdot \int_0^1 u \eta(u) dF_u = 0 \end{aligned} \quad (15)$$

Based on the revenue function, Equation (14) and (15), we have

$$\begin{aligned} \frac{dH(\epsilon)}{d\epsilon} \Big|_{\epsilon=0} &= \frac{dT(h(u) + \epsilon\eta(u))}{d\epsilon} \Big|_{\epsilon=0} \\ &= \frac{1}{\beta} \int_0^1 u \eta(u) w(u) dF_u \cdot e^{-s(h(u))} = 0 \end{aligned}$$

where

$$w(u) = (h(u))^{\frac{1}{\beta}-1} [1 - (\beta + 1)h(u)K^\beta(s(h(u)))]. \quad (16)$$

Thus, we have  $\int_0^1 u \eta(u) w(u) dF_u = 0$  if  $\int_0^1 u \eta(u) dF_u = 0$ .

We make a substitution  $\xi(u) = \alpha u^\alpha \eta(u)$ . Then for any function  $\xi(u)$  satisfying  $\int_0^1 \xi(u) du = 0$ , we have  $\int_0^1 \xi(u) w(u) du = 0$ . Based on this condition, we next prove

that  $w(u)$  ( $u \in (0, 1)$ ) is a constant function by contradiction. Because we assumed that  $\tilde{p}_I(u)$  is a continuous function on  $[0, 1]$ ,  $w(u)$  is also continuous on  $[0, 1]$ . Further, it must be uniformly continuous on  $[0, 1]$ . Suppose there exists  $u_1, u_2 \in (0, 1)$  satisfying that  $|w(u_1) - w(u_2)| > \epsilon$ . By the uniform continuity, there must exist  $\delta$  satisfying  $\forall |u' - u| < \delta$  and  $|w(u') - w(u)| < \frac{\epsilon}{3}$ . We conduct the function  $\xi(u)$  as follows:

$$\xi(u) = \begin{cases} 1 & \text{if } |u - u_1| < \delta, \\ -1 & \text{if } |u - u_2| < \delta, \\ 0 & \text{otherwise.} \end{cases}$$

It is obvious that  $\int_0^1 \xi(u) du = 0$ . However,

$$\begin{aligned} \left| \int_0^1 \xi(u) w(u) du \right| &= \left| \int_{u_1-\delta}^{u_1+\delta} w(u) du - \int_{u_2-\delta}^{u_2+\delta} w(u) du \right| \\ &\geq \left( |w(u_1) - w(u_2)| - \frac{2\epsilon}{3} \right) \cdot 2\delta \geq \frac{2\epsilon\delta}{3} > 0. \end{aligned}$$

This is contradictory with  $\int_0^1 \xi(u) w(u) du = 0$  under condition  $\int_0^1 \xi(u) du = 0$ . Thus,  $w(u)$  ( $u \in (0, 1)$ ) is a constant function. Further, based on Equation (16),  $h(u)$  is a constant function and  $\tilde{p}_I(u)$  is also a constant function. We denote  $\tilde{p}_I \triangleq \tilde{p}_I(u)$  for any  $u \in (0, 1)$  and then  $(0, 0, \tilde{p}_I)$  must be a revenue-optimal fixed pay-as-you-go pricing. Conversely, if  $(0, 0, \tilde{p}_I)$  is a revenue-optimal fixed pay-as-you-go pricing,  $(0, 0, \tilde{p}_I(u))$  must be a revenue-optimal demand-based pay-as-you-go pricing where  $\tilde{p}_I(u) = \tilde{p}_I$  for any  $u \in [0, 1] = [0, U]$ . Therefore, Proposition 4 concludes. ■

**Proof of Theorem 4:** This theorem could be proofed by the same method with Theorem 3. In particular, we could derive results which are similar with Proposition 3 and 4 to show that a welfare-optimal fixed pay-as-you-go pricing is also a welfare-optimal demand-based pay-as-you-go pricing, which generates no lower revenue compared with any demand-based two-part pricing. Thus, any welfare-optimal pay-as-you-go pricing is also a global welfare-optimal two-part pricing. ■

## REFERENCES

- [1] X. Wang, R. T. B. Ma, and Y. Xu, "The role of data cap in optimal two-part network pricing," *Proceedings of the 24th International World Wide Web Conference (WWW)*, 2015.
- [2] X. Wang, R. T. B. Ma, and Y. Xu, "The role of data cap in two-part pricing under market competition," *Proceedings of the 4th Smart Data Pricing Workshop (SDP)*, 2015.
- [3] X. Wang, R. T. B. Ma, and Y. Xu, "The role of data cap in two-part pricing under market competition," *IEEE Network*, vol. 30(2), 2016.
- [4] A. Odlyzko, "Internet pricing and the history of communications," *Computer networks*, vol. 36, no. 5, pp. 493–517, 2001.
- [5] C. Labovitz, D. McPherson, S. Iekel-Johnson, J. Oberheide, and F. Jahani, "Internet inter-domain traffic," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 75–86, 2011.
- [6] P. Nabipay, A. Odlyzko, and Z.-L. Zhang, "Flat versus metered rates, bundling, and 'bandwidth hog'," *6th Workshop on the Economics of Networks, Systems, and Computation*, 2011.
- [7] M. Morgan, "Pricing schemes key in LTE future," *Telecomasia. net. September 12*, 2011.
- [8] L. Segall, "Verizon ends unlimited data plan," *CNN Money. July*, vol. 6, 2011.
- [9] P. Taylor, "AT&T imposes usage caps on fixed-line broadband," *Financial Times. March*, vol. 14, 2011.

- [10] R. Edell and P. Varaiya, "Providing Internet access: What we learn from INDEX," *IEEE Network*, vol. 13, no. 5, pp. 18–25, 1999.
- [11] W. Dai and S. Jordan, "Design and impact of data caps," *IEEE Global Communications Conference*, pp. 1650–1656, 2013.
- [12] F. O. I. A. Committee, "Policy issues in data caps and usage-based pricing," *Annual Report*, 2013.
- [13] A. Schatz and S. E. Ante, "FCC chief backs usage-based broadband pricing," *Wall Street Journal*, December 2, 2010.
- [14] W. Y. Oi, "A disneyland dilemma: Two-part tariffs for a mickey mouse monopoly," *The Quarterly Journal of Economics*, pp. 77–96, 1971.
- [15] P. S. Calem and D. F. Spulber, "Multiproduct two part tariffs," *International Journal of Industrial Organization*, vol. 2, no. 2, pp. 105–115, 1984.
- [16] S. C. Littlechild, "Two-part tariffs and consumption externalities," *The Bell Journal of Economics*, pp. 661–670, 1975.
- [17] S. Scotchmer, "Two-tier pricing of shared facilities in a free-entry equilibrium," *The Rand Journal of Economics*, pp. 456–472, 1985.
- [18] W. Dai and S. Jordan, "How do ISP data caps affect subscribers?," *Telecommunications Policy Research Conference TPRC*, 2013.
- [19] A. Odlyzko, B. S. Arnaud, E. Stallman, and M. Weinberg, "Know your limits: Considering the role of data caps and usage based billing in Internet access service," *Public Knowledge*, April 23, 2012.
- [20] M. Chetty, R. Banks, A. Brush, J. Donner, and R. Grinter, "You're capped: understanding the effects of bandwidth caps on broadband use in the home," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3021–3030, 2012.
- [21] K. Poularakis, I. Pefkianakis, J. Chandrashekar, and L. Tassiulas, "Pricing the last mile: Data capping for residential broadband," *ACM SIGCOMM CoNEXT*, pp. 295–306, 2014.
- [22] P. Hande, M. Chiang, R. Calderbank, and J. Zhang, "Pricing under constraints in access networks: Revenue maximization and congestion management," *Proceedings of IEEE INFOCOM*, pp. 1–9, 2010.
- [23] S. Li, J. Huang, and S.-Y. Li, "Revenue maximization for communication networks with usage-based pricing," *Global Telecommunications Conference*, pp. 1–6, 2009.
- [24] T. Basar and R. Srikant, "Revenue-maximizing pricing and capacity expansion in a many-users regime," *Proceedings of IEEE INFOCOM*, pp. 294–301, 2002.
- [25] H. Shen and T. Basar, "Optimal nonlinear pricing for a monopolistic network service provider with complete and incomplete information," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 6, pp. 1216–1223, 2007.
- [26] P. Chander and L. Leruth, "The optimal product mix for a monopolist in the presence of congestion effects: A model and some results," *International Journal of Industrial Organization*, vol. 7, no. 4, pp. 437–449, 1989.
- [27] D. Reitman, "Endogenous quality differentiation in congested markets," *The Journal of Industrial Economics*, vol. 39, no. 6, pp. 621–647, 1991.
- [28] R. T. B. Ma, "Pay-as-you-go pricing and competition in congested network service markets," *Proceedings of the 22nd IEEE International Conference on Network Protocols (ICNP)*, pp. 257–268, 2014.
- [29] R. T. B. Ma and V. Misra, "The public option: a non-regulatory alternative to network neutrality," *IEEE/ACM Transactions on Networking*, vol. 21, pp. 1866–1879, Dec 2013.
- [30] T. Gryta, "FCC votes to allow municipal broadband, overruling two states' laws," *Wall Street Journal*, Feb 26, 2015.
- [31] R. T. B. Ma, J. C. Lui, and V. Misra, "On the evolution of the Internet economic ecosystem," *Proceedings of the 22nd International World Wide Web Conference*, pp. 849–860, 2013.
- [32] C.-K. Chau, Q. Wang, and D.-M. Chiu, "On the viability of Paris Metro pricing for communication and service networks," *Proceedings of IEEE INFOCOM*, pp. 1–9, 2010.
- [33] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2490–2498, 2000.
- [34] R. Jain, T. Mullen, and R. Hausman, "Analysis of Paris Metro pricing strategy for QoS with a single service provider," *Quality of Service IWQoS*, pp. 44–58, 2001.